On Generalized Processor Sharing and Objective Functions: Analytical Framework

Jasper Vanlerberghe[1], Joris Walraevens[1], Tom Maertens[1], Stijn De Vuyst[2], and Herwig Bruneel[1]

[1] Stochastic Modelling and Analysis of Communication Systems Research Group
Department of Telecommunications and Information Processing (TELIN)
Ghent University (UGent)
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium
{jpvlerbe,jw,tmaerten,hb}@telin.UGent.be

[2] Supply Networks & Logistics Research Center (SNLRC)
Department of Industrial Management
Ghent University (UGent)
Technologiepark 903, B-9052 Zwijnaarde, Belgium
Stijn.DeVuyst@UGent.be

Abstract. Today, telecommunication networks host a wide range of heterogeneous services. Some demand strict delay minima, while others only need a best-effort kind of service. To achieve service differentiation, network traffic is partitioned in several classes which is then transmitted according to a flexible and fair scheduling mechanism. Telecommunication networks can, for instance, use an implementation of Generalized Processor Sharing (GPS) in its internal nodes to supply an adequate Quality of Service to each class. GPS is flexible and fair, but also notoriously hard to study analytically. As a result, one has to resort to simulation or approximation techniques to optimize GPS for some given objective function. In this paper, we set up an analytical framework for two-class discrete-time probabilistic GPS which allows to optimize the scheduling for a generic objective function in terms of the mean unfinished work of both classes without the need for exact results or estimations/approximations for these performance characteristics. This framework is based on results of strict priority scheduling, which can be regarded as a special case of GPS, and some specific unfinished-work properties in two-class GPS. We also apply our framework on a popular type of objective functions, i.e., convex combinations of functions of the mean unfinished work. Lastly, we incorporate the framework in an algorithm to yield a faster and less computation-intensive result for the optimum of an objective function.

Keywords: Generalized Processor Sharing (GPS), optimization, queueing, scheduling, objective function

1 Introduction

Times when telecommunication networks were used for one single service like telephony or television are long gone. Nowadays, telecommunication systems host

a wide collection of services. Amongst those services are the traditional services like internet, telephony, and television, but modern telecommunication networks also support more demanding interactive multimedia services such as online gaming and video conferencing. Every service desires other network requirements in order to deliver a certain Quality of Service (QoS) or Quality of Experience (QoE) to the end user [4, 12]. Hence, the network needs a way to differentiate services. This can be achieved by dividing the network traffic into several classes and implementing some kind of priority scheduling amongst those classes.

Giving strict priority to the different classes in a hierarchical way may not be flexible enough. Additionally, strict priority is not fair since a high load of a higher-priority class can lead to starvation of lower-priority classes [3, 5, 6]. One scheduling mechanism able to deliver fairness and flexibility is Generalized Processor Sharing (GPS) [9, 10]. With GPS, each class is given a certain weight and the available link capacity is shared according to the weights of the backlogged classes. In this way, no capacity goes to waste and each class gets a minimum capacity. Starvation is thus not an issue for GPS. The biggest drawback of all GPS-like scheduling mechanisms is the complexity of obtaining analytical results for their performance characteristics, such as (mean) delays, queue contents or unfinished work. As a consequence, it is hard to analytically determine the optimal weights minimizing an objective function that depends on these characteristics.

In this paper, we consider a discrete-time, probabilistic emulation of single-server, two-class GPS. The weights of the two classes are normalized such that they sum up to one. Then setting the weight of one of the classes to 1 implies that this class has strict priority over the other class; so strict priority scheduling can be seen as a special case of GPS. Strict priority scheduling is well-studied and allows, in many important cases, for an explicit analytical solution. In the remainder, we first show how to use (i) results of strict priority scheduling and (ii) some specific properties of the unfinished work in this GPS system, to transform a generic objective function as to determine its behaviour. This transformation leads to a format which does not require exact results or estimations/approximations for performance characteristics. It allows to find the number of (local) extrema and inflection points and determine the values of the objective function in these points. This analytical framework can thus be invaluable to network operators as it provides an opportunity for them to quickly estimate the need for the possible time-consuming quest for the optimal weight. Furthermore, in case such a quest is recommended, the framework helps to make it more efficient.

Secondly, we extend the results of [13], where we have considered the same GPS-like model. Our theoretical study there concentrated on the behaviour of a specific type of objective functions, i.e., convex combinations of increasing functions (both convex, concave or linear) of the mean unfinished work of both classes. Here, we consider a more general type of objective functions, i.e., convex combinations of (more) general functions of the mean unfinished work of both classes. This removes the requirement for the functions of the unfinished work to be increasing and both either convex, concave or linear in the relevant domain. We show how to study the behaviour of this popular type of objective functions, based on some new theorems and by using the analytical framework

Fig. 1. GPS system at hand

described in the first part of our paper. Finally, we show how to exploit our theoretical results in a sensitivity analysis on the objective function parameters of the optimum. This is very useful for network operators, as it allows them to avoid simulating the system or use complex approximation techniques.

In the next section, we unfold our analytical framework, for a generic objective function. We show some illustrative examples of our findings in Sect. 3. In Sect. 4, we apply this framework to study a more specific type of objective functions. We prove a theorem to easily carry out the higher order derivative test on this type of objective functions and demonstrate how to study the behaviour of these objective functions with respect to (w.r.t.) the GPS weights and the coefficients in these functions. Before summarizing the most important conclusions, we illustrate the gains of the framework with a practical example in Sect. 5.

2 Analytical Framework

As already mentioned, we consider a discrete-time GPS model with one server and two traffic classes. We denote these classes by class 1 and class 2. The weights of both classes are normalized: class 1 is assigned weight β and class 2 is assigned weight $1-\beta$, with $0 \leq \beta \leq 1$. So when both classes are backlogged, the server will choose a class-1 packet with probability β and a class-2 packet with probability $1-\beta$. In case one of the classes is not backlogged, the server picks a packet of the other class. For the sake of convenience, we assume that both classes have their own queue. This GPS system is depicted in Fig. 1. The cases $\beta = 0$ and $\beta = 1$ reduce to strict priority scheduling.

Next, we define the unfinished work in a queue at the beginning of a slot as the sum of the residual service/transmission times of the packets present in the queue at that moment. It is obvious that the unfinished work in both queues depends on the parameter β: the lower (higher) the value of β, the less capacity for class 1 (2). Therefore, the mean unfinished work in queue j at the beginning of a random slot in steady state is denoted as $\bar{w}_j(\beta)$ $(j = 1, 2)$. We assume that the arrival and service processes of the system at hand are such that $\bar{w}_1(\beta)$ and $\bar{w}_2(\beta)$ satisfy two important properties, i.e.,

Property 1. Function $\bar{w}_1(\beta) + \bar{w}_2(\beta)$ is independent of β.
Property 2. $\bar{w}_2(\beta)$ and $\bar{w}_1(\beta)$ are analytic and strictly monotonic (increasing and decreasing, respectively) w.r.t. β, i.e., on the interval $[0, 1]$.

These properties basically follow (i) from the observation that GPS is a work-conserving scheduling mechanism and (ii) that class 2 is given less capacity with increasing β, respectively (see, e.g., [13, 14, 18] for more formal proofs). (In fact,

only strict monotonicity of one of the $\bar{w}_j(\beta)$ functions is required as Property 1 implies the other is stricly monotone as well.) The first property states that the mean total unfinished work is a constant w.r.t. β. This constant, say $\bar{w}_T$, can be calculated explicitly for a whole range of arrival and service processes. Indeed, since the scheduling mechanism is work-conserving, in every time slot one unit of work (if any) is executed. So to study the total unfinished work, we can consider the system to be a single queue consisting of units of work which are, for instance, executed according to a First-In-First-Out (FIFO) scheduling. Discrete-time, single-queue systems with a FIFO scheduling are much easier to study analytically than multi-queue systems (see, e.g., [2]).

The second property, furthermore, has important consequences as well. Since $\bar{w}_2(\beta)$ is continuous on the interval $[0,1]$ and takes the values $\bar{w}_2(0)$ and $\bar{w}_2(1)$ at each end of that interval, we can apply the intermediate value theorem to conclude that $\bar{w}_2(\beta)$ takes any value between $\bar{w}_2(0)$ and $\bar{w}_2(1)$ at minimum one point within $[0,1]$. From the *strict* monotonic increasing property, finally, we have that $\bar{w}_2(\beta)$ is bijective on $[0,1]$, i.e., that there is a one-to-one correspondence between all values in $[0,1]$ and all values in $[\bar{w}_2(0), \bar{w}_2(1)]$. For $\bar{w}_1(\beta)$, we can set up a similar reasoning. However, since $\bar{w}_1(\beta)$ is decreasing w.r.t. β, the image of $\bar{w}_1(\beta)$ is the interval $[\bar{w}_1(1), \bar{w}_1(0)]$. For ease of notation, we define the intervals $[\bar{w}_1(1), \bar{w}_1(0)]$ and $[\bar{w}_2(0), \bar{w}_2(1)]$ as Ω_1 and Ω_2, respectively. For a whole range of arrival and service processes, Ω_1 and Ω_2 can be determined explicitly, as they arise from results of strict priority systems (see, e.g., [16,17]).

Now we turn to optimization. The optimal β is defined as the β-value that minimizes some objective function. In the context of scheduling mechanisms, objective functions are often constructed in terms of (mean) delays or holding times. Here, we assume the objective function to be a generic function of the mean unfinished work in both queues, i.e., $f(\bar{w}_1(\beta), \bar{w}_2(\beta))$. [3] It is clear that the objective function $f(\bar{w}_1(\beta), \bar{w}_2(\beta))$ can be seen as a function in terms of β, say $F(\beta) \triangleq f(\bar{w}_1(\beta), \bar{w}_2(\beta))$, with domain $[0,1]$. The objective function $f(\bar{w}_1(\beta), \bar{w}_2(\beta))$, however, can also be expressed in terms of another single variable. In particular, the first unfinished-work property states that $\bar{w}_1(\beta) = \bar{w}_T - \bar{w}_2(\beta)$, implying that $f(\bar{w}_1(\beta), \bar{w}_2(\beta))$ can be expressed in terms of $\bar{w}_2(\beta)$ only. With a slight abuse of notation, we define this format as $f^*(\bar{w}_2)$. It is obvious that $F = f^* \circ \bar{w}_2$. As mentioned earlier, analytical results for $\bar{w}_2(\beta)$ are notoriously complex to obtain, so we need estimations/approximations for this performance measure to study the objective function $f(\bar{w}_1(\beta), \bar{w}_2(\beta))$ in terms of β. However, since we know the image ($\Omega_2 = [\bar{w}_2(0), \bar{w}_2(1)]$) and behaviour (continuous and strictly increasing) of $\bar{w}_2(\beta)$, we can already study $f(\bar{w}_1(\beta), \bar{w}_2(\beta))$ in terms of $\bar{w}_2(\beta)$ instead of β, with domain Ω_2 instead of $[0,1]$ (i.e., studying $f^*(\bar{w}_2)$).

Consequently, we can observe the number of extrema and inflection points and determine the values of $f(\bar{w}_1(\beta), \bar{w}_2(\beta))$ in these points without running sim-

[3] In some specific cases, it is possible to find easy relations between the mean unfinished work in both queues and the corresponding mean queue contents and/or delays (see [13] for such a case). Then one can initially consider an objective function in terms of, e.g., mean packet delays, which may be practically most relevant in the context of heterogeneous services, translate it to an objective function in terms of the mean unfinished work in both queues, and apply our framework to study the latter.

ulations or relying on possibly inaccurate approximate expressions. Obviously, we do not know the β-values corresponding to these points (except when they coincide with the endpoints). To determine these β-values, we still need estimations/approximations. Now some preliminary conclusions can be drawn from the behaviour of $f^*(\bar{w}_2)$. For instance, the minimum can be in the endpoints $\beta = 0$ or $\beta = 1$. In that case, strict priority is optimal and we do not have to simulate. Another possible conclusion is that the difference in the objective function between the minimum and one of the endpoints is too small to justify a time-consuming quest for the β-value corresponding to the minimum. Summarized, from the analysis of $f^*(\bar{w}_2)$, an interval in Ω_2 with an acceptable value for the objective function can be selected. The optimization problem then reduces to finding a value of β for which the continuous and monotonic function $\bar{w}_2(\beta)$ reaches a value in the selected interval (stopping criterium). In the next section, we demonstrate these findings by means of some illustrative examples.

3 Some Illustrative Examples

For the examples, we consider one-slot service times and a two-dimensional binomial arrival process characterized by the joint probability generating function

$$A(z_1, z_2) \triangleq \left(1 + \frac{\lambda_1}{N}(z_1 - 1) + \frac{\lambda_2}{N}(z_2 - 1)\right)^N, \tag{1}$$

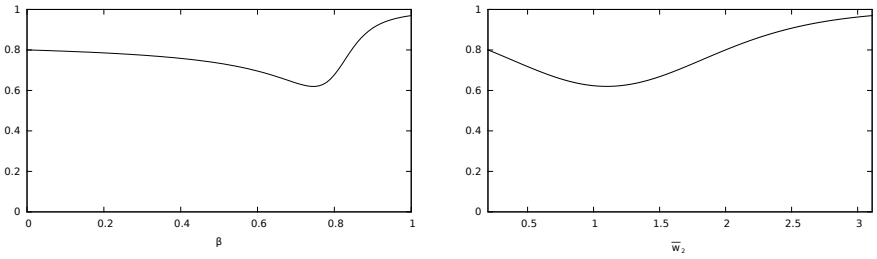
of the independently and identically distributed number of class-1 and class-2 arrivals in a slot. Here, λ_j $(j = 1, 2)$ is the arrival rate of class-j packets. This is the arrival process in a queue of an NxN output-queueing switch with Bernoulli arrivals at its inlets and with independent and uniform routing towards the outlets. Parameter N expresses the maximum total number of arrivals in a queue during a slot. For the sake of convenience, we also introduce the parameters λ_T and α, indicating the total arrival rate (i.e., $\lambda_T = \lambda_1 + \lambda_2$) and the fraction of class-1 packets in the overall arrival stream (i.e., $\alpha = \frac{\lambda_1}{\lambda_T}$), respectively. A queueing model with this type of arrival process, one-slot service times and strict priority scheduling is, for instance, studied in [15]. Adopting some concrete values for the parameters of the arrival process, the results of [15] can be used to calculate $\bar{w}_j(0)$ and $\bar{w}_j(1)$ $(j = 1, 2)$. For $N = 16$, $\lambda_T = 0.9$, and $\alpha = 0.8$, for example, we find that

$$\begin{aligned} \bar{w}_1(0) &= 4.50, \qquad \bar{w}_2(0) = 0.20, \\ \bar{w}_1(1) &= 1.59, \qquad \bar{w}_2(1) = 3.11. \end{aligned} \tag{2}$$

Then $\bar{w}_T = \bar{w}_1(\cdot) + \bar{w}_2(\cdot) = 4.70$ and the intervals Ω_1 and Ω_2, defined in the previous section, equal $[1.59, 4.5]$ and $[0.2, 3.11]$, respectively.

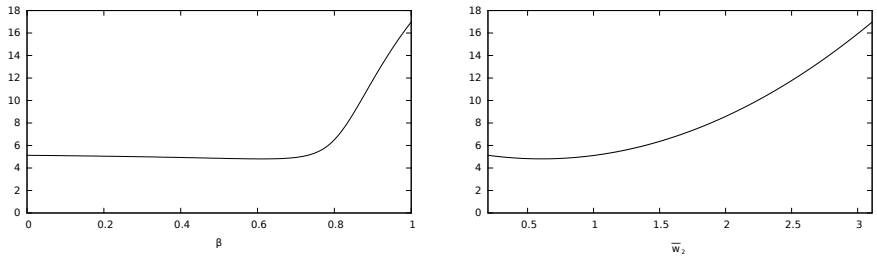
As objective function for our first example, we add two logistic functions:

$$f_1(\bar{w}_1(\beta), \bar{w}_2(\beta)) \triangleq \frac{1}{1 + e^{-2\bar{w}_1(\beta)+8}} + \frac{1}{1 + e^{-2\bar{w}_2(\beta)+3}}. \tag{3}$$



(a) $f_1(\bar{w}_1(\beta), \bar{w}_2(\beta))$ (b) $f_1^*(\bar{w}_2)$

Fig. 2. Comparison between the objective function $f_1(\bar{w}_1(\beta), \bar{w}_2(\beta))$ and $f_1^*(\bar{w}_2)$



(a) $f_2(\bar{w}_1(\beta), \bar{w}_2(\beta))$ (b) $f_2^*(\bar{w}_2)$

Fig. 3. Comparison between the objective function $f_2(\bar{w}_1(\beta), \bar{w}_2(\beta))$ and $f_2^*(\bar{w}_2)$

Applying the framework of the previous section and using the values of (2) then yields $f_1(\bar{w}_1(\beta), \bar{w}_2(\beta))$ in terms of $\bar{w}_2(\beta)$ only, i.e.,

$$f_1^*(\bar{w}_2) = \frac{1}{1 + e^{2\bar{w}_2 - 1.4}} + \frac{1}{1 + e^{-2\bar{w}_2 + 3}}. \tag{4}$$

This function is plotted in Fig. 2b. In Fig. 2a, $f_1(\bar{w}_1(\beta), \bar{w}_2(\beta))$ is plotted as a function of β; this figure results from a Monte-Carlo simulation over one million slots (see further for more details), as it is notoriously complex to find analytical results for $\bar{w}_2(\beta)$. We see from the figures that both graphs have equal characteristics. Both graphs, for instance, have the same number of minima (i.e., one). Also, the ranges of both graphs are the same. So, $f_1^*(\bar{w}_2)$ helps identifying the minimum value of the objective function and determining how much this value differs from the values in the endpoints. As opposed to $f_1(\bar{w}_1(\beta), \bar{w}_2(\beta))$, however, we can plot $f_1^*(\bar{w}_2)$ right away.

Nevertheless, we cannot conclude from the behaviour of $f_1^*(\bar{w}_2)$ at what β-value this minimum occurs (say $\beta_{\min}$). We only know that $\bar{w}_2(\beta_{\min}) = 1.1$ at $F = 0.62$. So for instance if we are satisfied with the objective function within 2% of its minimum (i.e. F smaller than 0.6324), we then argue that we need $\bar{w}_2$ to be in the interval $[0.9, 1.3]$. This subsequently is the stopping criterium for a simulation or approximation procedure on the function $\bar{w}_2(\beta)$. As can be seen from Fig. 2a, the procedure should result in a β_{opt} value in the interval $[0.71, 0.77]$, as to have a value for the objective function within 2% of the minimum.

As a second example, we examine the objective function

$$f_2(\bar{w}_1(\beta), \bar{w}_2(\beta)) \triangleq (0.5\bar{w}_1(\beta))^2 + (1.3\bar{w}_2(\beta))^2. \quad (5)$$

Applying the framework of Sect. 2 and using the values in (2) then leads to

$$f_2^*(\bar{w}_2) = (2.35 - 0.5\bar{w}_2)^2 + (1.3\bar{w}_2)^2. \quad (6)$$

We know from the theory of [13] that $f_2(\bar{w}_1(\beta), \bar{w}_2(\beta))$ will have a minimum different from the endpoints $\beta = 0$ or $\beta = 1$, because of the convex character of the squaring in the objective function. Plots for this second example are found in Fig. 3. Function $f_2^*(\bar{w}_2)$ provides us with the value of the objective function in its minimum (see Fig. 3b). The difference between the value of the objective function at $\beta = 0$ (5.13) and the minimum value (4.81) is perhaps not worth the effort to search for $\beta_{\min}$ (the difference is 6.6%). Fig. 3b can thus be used in advance to decide whether significant gain can be won by searching $\beta_{\min}$ compared to using the priority cases $\beta = 0$ or $\beta = 1$.

If, on the other hand, we make the same reasoning as in the previous example, thus allowing at most 2% deviation of the minimum, we need the objective function F to be smaller than 4.90 and consequently $\bar{w}_2$ in $[0.38, 0.83]$. An accompanying simulation/approximation method should then result in a β_{opt}-value in the interval $[0.44, 0.69]$.

The reasoning done here can be done for an arbitrarily complex objective function in $\bar{w}_j(\beta)$. Indeed, every objective function minimization problem can be translated to a problem of finding the corresponding β for a certain performance vector (with or without some error margin). This effectively simplifies procedures; we will come back to this in Sect. 5.

Note: Critical for this method is the availability of the values for $\bar{w}_j(0)$ and $\bar{w}_j(1)$ $(j = 1, 2)$. However, it does not matter how these values are obtained. For strict priority systems, a lot of analytical results are available; this is, for instance, the case for the arrival and service processes we have used in the examples above (see [15]). For more complex arrival and/or service processes, this is not necessarily the case. To still obtain accurate values for $\bar{w}_j(0)$ and $\bar{w}_j(1)$ $(j = 1, 2)$, we can, for example, simulate the system for $\beta = 0$ and $\beta = 1$ only.

4 Framework Application

In this section, we address another important issue, namely the selection of an objective function and in particular the influence of this selection on the optimum. In a first step, a network operator chooses the type of relation of each performance characteristic in the objective function. When equal increments for high or low values should have an equal influence on the objective function, a linear relation can be used. The behaviour of other types of relations can easily be derived from a plot of the corresponding function. Other examples are a squared relation (as in f_2, see (5)) or a logistic one (as in f_1, see (3)). This choice of relation is closely related to the relation between QoS and QoE [4] and the choice of utility functions [7,8]. A second question for the operator is how to

weigh the performance of both classes. An answer to this question is less clear and in most cases more arbitrarily chosen. In the next paragraphs, we derive a method the operator can use to do a sensitivity analysis on these weights. This way, he can assess the impact of his choice on the behaviour and the resulting minimum of the objective function.

Following the reasoning above, we propose the following template for the objective function:

$$F(\gamma, \beta) = \gamma g_1(\bar{w}_1(\beta)) + (1-\gamma) g_2(\bar{w}_2(\beta)), \tag{7}$$

whereby other characteristics can be achieved by incorporating them in g_j as noted before. The parameter γ, assumed to be in the interval $[0, 1]$, serves as weight parameter, on which we want to do our sensitivity analysis. When $\gamma = 0$, the objective function only takes into account $\bar{w}_2(\beta)$; when $\gamma = 1$, only $\bar{w}_1(\beta)$ plays a role.

The exposition in the previous sections allows us to study this objective function for a specific value of γ. In [13], we already studied a subclass of this kind of objective functions. The analysis there was limited to increasing g_j that were either both linear, convex or concave. We proved that for linear or concave g_j the optimal β-value is always one of the endpoints (i.e., $\beta = 0$ or $\beta = 1$). For convex g_j, on the other hand, we have found that for certain values of γ the objective function reaches a minimum for some β different from 0 and 1.

In [13], we were able to study the objective function for all γ in one effort, but we were limited to increasing functions g_j. Now we generalize the results of [13], allowing all functions g_j. The function

$$\phi(\beta) \triangleq \frac{g_2'(\bar{w}_2(\beta))}{g_2'(\bar{w}_2(\beta)) + g_1'(\bar{w}_1(\beta))}, \tag{8}$$

defined in [13], plays a key role.

Theorem 1. *Assume g_j is continuously differentiable in Ω_j. Then $\gamma = \phi(\beta)$ if and only if (iff) $\frac{\partial F}{\partial \beta}(\gamma, \beta) = 0$.* [4]

Proof. From (7), we find that $\frac{\partial F}{\partial \beta}(\gamma, \beta) = 0$ is equivalent with

$$[g_2'(\bar{w}_2(\beta)) - \gamma(g_1'(\bar{w}_1(\beta)) + g_2'(\bar{w}_2(\beta)))]\bar{w}_2'(\beta) = 0, \tag{9}$$

where we have used that $\bar{w}_1'(\beta) = -\bar{w}_2'(\beta)$, see Property 1. According to Property 2 ($\bar{w}_2'(\beta) > 0$) and (9), this leads to

$$\gamma = \phi(\beta). \tag{10}$$

β-values for which $\phi(\beta) = \gamma$ are the critical points of the objective function $F(\gamma, \beta)$. Critical points can be either extrema or inflection points with a horizontal asymptote.

For simplicity, we assume $\phi(\beta)$ to be continuous. Discontinuities only occur for β-values for which $g_2'(\bar{w}_2(\beta)) = -g_1'(\bar{w}_1(\beta))$. For these β-values,

$$\frac{\partial F}{\partial \beta}(\gamma, \beta) = g_2'(\bar{w}_2(\beta))\bar{w}_2'(\beta). \tag{11}$$

[4] In fact, Theorem 1 is a generalization of Lemma 1 in [13].

As $\bar{w}_2'(\beta)$ is positive, the objective function will increase or decrease like g_2. We will disregard these cases in the remainder, as the discontinuities in $\phi(\beta)$ do not lead to special cases for $F(\gamma, \beta)$. The extensions are straightforward but only result in more involved expressions.

To be able to distinguish extrema from inflection points, we need higher-order derivatives of the objective function. These will allow us to perform the higher-order derivative test on the objective function. Therefore, we extend Theorem 1.

Theorem 2. *Assume g_j is n times continuously differentiable in Ω_j. Then $\phi^{(i-1)}(\beta) = 0, \forall i = 2, ..., n$, and $\phi(\beta) = \gamma$ iff $\frac{\partial^i F}{\partial \beta^i}(\phi(\beta), \beta) = 0, \forall i = 1, ..., n$. Here, $\phi^{(j)}(\beta)$ denotes the j-th derivative of $\phi(\beta)$.*

Proof. (by induction) The base case $n = 1$ follows from Theorem 1. For the induction hypothesis, assume that $\phi(\beta) = \gamma$ and $\phi^{(i-1)}(\beta) = 0$ for $i = 2, ..., n-1$ iff $\frac{\partial^i F}{\partial \beta^i}(\phi(\beta), \beta) = 0$ for $i = 1, ..., n-1$. To complete the theorem, we prove that $\phi^{(n-1)}(\beta) = 0$ iff $\frac{\partial^n F}{\partial \beta^n}(\phi(\beta), \beta) = 0$. We find that

$$\frac{\partial^n F}{\partial \beta^n} = \frac{\partial^{n-1}}{\partial \beta^{n-1}} \left(\frac{\partial F}{\partial \beta} \right) = \frac{\partial^{n-1}}{\partial \beta^{n-1}} \Big(\bar{w}_2'(\beta) \big(g_1'(\bar{w}_1(\beta)) + g_2'(\bar{w}_2(\beta)) \big) \big(\phi(\beta) - \gamma \big) \Big),$$

where we have used (9). Define, furthermore, $\Delta(\gamma, \beta)$ as $\phi(\beta) - \gamma$ and $\chi(\beta)$ as $\bar{w}_2'(\beta) \big(g_1'(\bar{w}_1(\beta)) + g_2'(\bar{w}_2(\beta)) \big)$. Then we can write

$$\frac{\partial^n F}{\partial \beta^n} = \frac{\partial^{n-1}}{\partial \beta^{n-1}} \Big(\chi(\beta) \Delta(\gamma, \beta) \Big) = \sum_{r=0}^{n-1} \binom{n-1}{r} \chi^{(n-1-r)}(\beta) \frac{\partial^r \Delta}{\partial \beta^r}(\gamma, \beta). \tag{12}$$

We know from the induction hypothesis that $\frac{\partial^r \Delta}{\partial \beta^r}(\gamma, \beta) = \frac{\partial^r (\phi(\beta) - \gamma)}{\partial \beta^r} = \frac{\partial^r \phi(\beta)}{\partial \beta^r} = \phi^{(r)}(\beta) = 0$, $r = 1, ..., n-2$, and that $\Delta(\gamma, \beta) = 0$. This yields

$$\frac{\partial^n F}{\partial \beta^n} = \binom{n-1}{n-1} \chi^{(0)}(\beta) \frac{\partial^{n-1} \Delta}{\partial^{n-1} \beta}(\gamma, \beta) = \bar{w}_2'(\beta) \big(g_1'(\bar{w}_1(\beta)) + g_2'(\bar{w}_2(\beta)) \big) \phi^{(n-1)}(\beta).$$

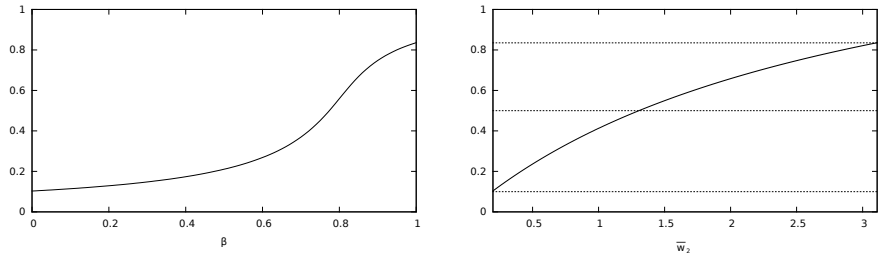
Strict monotonicity of $\bar{w}_2(\beta)$ and the continuity assumption of $\phi(\beta)$ that we made earlier prove that $\phi^{(n-1)}(\beta) = 0$ iff $\frac{\partial^n F}{\partial \beta^n}(\phi(\beta), \beta) = 0$.

The next corollary follows directly from Theorem 2:

Corollary 1. *If $\gamma = \phi(\beta)$, $\phi^{(1)}(\beta) = \cdots = \phi^{(n)}(\beta) = 0$ and $\phi^{(n+1)}(\beta) \neq 0$, then $F(\gamma, \beta)$ has a local extremum at β if n is even and an inflection point at β if n is odd.*

With this corollary, we can determine the behaviour of $F(\gamma, \beta)$ by studying the behaviour of $\phi(\beta)$. Suppose, for instance, that $\phi(\beta)$ has one inflection point $\hat{\beta}$ with horizontal tangent. Then $\phi^{(2)}(\hat{\beta}) = \phi^{(1)}(\hat{\beta}) = 0$ and $\phi^{(3)}(\hat{\beta}) \neq 0$, and, as a consequence, $F(\gamma, \beta)$ has an extremum at $\hat{\beta}$ when $\gamma = \phi(\hat{\beta})$.

Unfortunately, we do not have a formula for $\phi(\beta)$, as we do not have explicit analytical results for the functions $\bar{w}_j(\beta)$. This is where the framework of Sect. 2 comes into play. In particular, the function $\phi(\beta)$ can be translated into a function



(a) $\phi(\beta)$ (b) $\phi^*(\bar{w}_2)$

Fig. 4. For F_3: Comparison between $\phi(\beta)$ and $\phi^*(\bar{w}_2)$

in terms of $\bar{w}_2(\beta)$ instead of β (i.e. $\phi^*(\bar{w}_2)$ in the remainder). As there is a one-to-one mapping between the values in Ω_2 and the values in $[0, 1]$, we find that

$$\phi^*(\bar{w}_2) = \frac{g_2'(\bar{w}_2)}{g_2'(\bar{w}_2) + g_1'(\bar{w}_T - \bar{w}_2)}. \tag{13}$$

Using the framework, the previous corollary is reformulated as follows:

Corollary 2. *If* $\gamma = \phi^*(\bar{w}_2)$, $\phi^{*(1)}(\bar{w}_2) = \cdots = \phi^{*(n)}(\bar{w}_2) = 0$ *and* $\phi^{*(n+1)}(\bar{w}_2) \neq 0$, *then* $F^*(\gamma, \bar{w}_2)$ *has a local extremum at* $\bar{w}_2$ *if* n *is even and an inflection point at* $\bar{w}_2$ *if* n *is odd. As* $\bar{w}_2(\beta)$ *is bijective on* $[0, 1]$, $F(\gamma, \beta)$ *also has a local extremum at* β *if* n *is even and an inflection point at* β *if* n *is odd.*

Hereby, we defined $F^*(\gamma, \bar{w}_2)$ analogously to the other functions marked with a star, using the framework presented in Sect. 2.

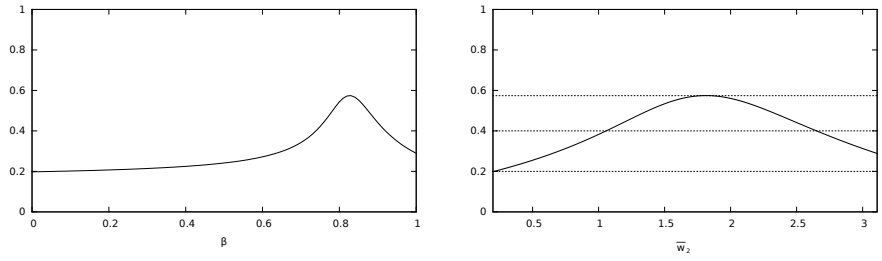
As in Sect. 3, we have composed figures to compare $\phi(\beta)$ (see Fig. 4a) with $\phi^*(\bar{w}_2)$ (see Fig. 4b). For these figures, we have used the objective function

$$F_3(\gamma, \beta) = \gamma(0.5\sqrt{2}\bar{w}_1(\beta))^2 + (1-\gamma)(1.3\sqrt{2}\bar{w}_2(\beta))^2. \tag{14}$$

and the same arrival and service processes as in Sect. 3. It should be noticed that $F_3(0.5, \beta) = f_2(\bar{w}_1(\beta), \bar{w}_2(\beta))$ (see (5)); we refer to this equivalence later. We can draw similar conclusions as in the previous section. For instance, we can see that both graphs have the same range.

Using the aforementioned corollary, we see that for $\gamma \in [0.1, 0.83]$ (obtained using strict priority scheduling results only, see [13]), $F_3(\gamma, \beta)$ reaches an extremum at a β-value different from 0 or 1. In particular for $\gamma \in [0.1, 0.83]$, there is a β-value and corresponding $\bar{w}_2(\beta)$-value, say $\hat{\beta}$ and $\bar{w}_2(\hat{\beta})$, respectively, for which $\gamma = \phi(\beta) = \phi^*(\bar{w}_2(\beta))$. Visually, this $\hat{\beta}$ and $\bar{w}_2(\hat{\beta})$ can be presented in the Cartesian coordinate systems $(\beta, \phi(\beta))$ and $(\bar{w}_2, \phi^*(\bar{w}_2))$, by drawing a horizontal line at the chosen value of γ (see Fig. 4b); the intersection points of the horizontal lines and curves of $\phi(\beta)$ and $\phi^*(\bar{w}_2)$ then yield $\hat{\beta}$ and $\bar{w}_2(\hat{\beta})$, respectively.

Now according to Theorem 2, $\frac{\partial F_3}{\partial \beta}(\gamma, \hat{\beta}) = 0$ and we have an extremum at $\hat{\beta}$ if $\frac{\partial^2 F_3}{\partial \beta^2}(\gamma, \hat{\beta}) \neq 0$ or, equivalently, if $\phi^{(1)}(\hat{\beta}) \neq 0$. From Fig. 3, where we have depicted $F_3(\gamma, \beta)$ for $\gamma = 0.5$, we can see that $F_3(\gamma, \beta)$ is decreasing at $\beta = 0$ for $\gamma \in [0.1, 0.83]$. As we have a $\hat{\beta}$ for which $\gamma = \phi(\beta)$ in that interval, $F_3(\gamma, \beta)$ has an extremum, which is necessarily a minimum. Summarized, the couples



(a) $\phi(\beta)$ (b) $\phi^*(\bar{w}_2)$

Fig. 5. For F_4: Comparison between $\phi(\beta)$ and $\phi^*(\bar{w}_2)$

$(\phi(\beta), \beta)$, indicated by the curve in the figure, are parameter combinations for (γ, β) that minimize the objective function. For $\gamma < 0.1$, there is no β for which $\gamma = \phi(\beta)$ and thus, according to Theorem 1, $F_3(\gamma, \beta)$ has no extremum between 0 and 1. Since $F_3(\gamma, \beta)$ is increasing at $\beta = 0$, $F_3(\gamma, \beta)$ is increasing w.r.t. all β. Analogously, for $\gamma > 0.83$, the objective function is decreasing w.r.t β.

As a last example, we look at the objective function

$$F_4(\gamma, \beta) = \frac{\gamma}{1 + e^{-3\bar{w}_1(\beta)+9}} + \frac{1-\gamma}{1 + e^{-4\bar{w}_2(\beta)+7}}. \tag{15}$$

Using the same arrival and service processes and the same arrival process parameters as before, we depict the corresponding $\phi(\beta)$ in Fig. 5a and $\phi^*(\bar{w}_2)$ in Fig. 5b. Remember that the former is obtained via simulations, while the latter can be drawn directly. We can make the same reasoning as before. $\phi(\beta)$ and $\phi^*(\bar{w}_2)$ will have intersection points with horizontal lines at $\gamma \in [0.2, 0.58]$ only. For a γ-value in this interval, Theorem 1 dictates that $F_4(\gamma, \beta)$ will have extrema or inflection points. Using the corollary of Theorem 2, we know that inflection points only occur when also $\phi'(\beta) = 0$, so when $\phi(\beta)$ and $\phi^*(\bar{w}_2)$ have an extremum. For the example at hand, this occurs for $\gamma = 0.58$ and $\bar{w}_2 = 1.8$ (and from simulation, $\beta = 0.82$).

At $\bar{w}_2 = 0.2$ ($\beta = 0$), $F_4^*(\gamma, \bar{w}_2)$ is decreasing if $\gamma > 0.2$ and increasing if $\gamma < 0.2$ (this can easily be seen from (15)). So if we take $\gamma = 0.4$, we find that $F_4^*(\gamma, \bar{w}_2)$ is decreasing at $\bar{w}_2 = 0.2$. Furthermore, $F_4^*(0.4, \bar{w}_2)$ reaches an extremum at $\bar{w}_2 = 1.1$ because at that $\bar{w}_2$-value $\phi^*(\bar{w}_2)$ intersects with a horizontal line at 0.4 (see Fig. 5b). This extremum is necessarily a minimum. For higher values of $\bar{w}_2$, $F_4^*(\gamma, \bar{w}_2)$ is increasing again. At $\bar{w}_2 = 2.7$, we once more have a point of intersection between $\phi^*(\bar{w}_2)$ and the horizontal line at 0.4, and, hence, $F_4^*(\gamma, \bar{w}_2)$ has a second extremum, in this case a maximum. These conclusions can be verified in Fig. 6, where we have plotted $F_4^*(0.4, \bar{w}_2)$.

Using similar arguments for other values of γ, we constructed an annotated version of Fig. 5a in Fig. 7. In this figure, the arrows indicate the behaviour of $F_4(\gamma, \beta)$ for the (γ, β)-values in that area. We see that $F_4(\gamma, \beta)$ is decreasing above the curve of $\phi(\beta)$ and increasing under the curve. In fact, if we draw a path in the unit square (a collection of couples for (γ, β)), we know that the behaviour of the objective function is indicated by the arrows in the figure. Furthermore, the sign (and thus the behaviour) of $\frac{\partial F_4}{\partial \beta}(\gamma, \beta)$ will only change if the path

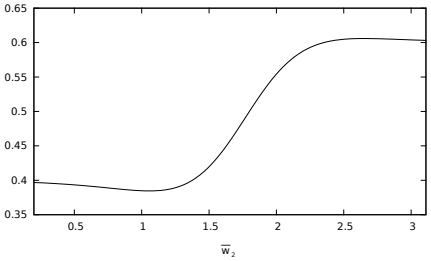


Fig. 6. Objective function $F_4^*(\gamma, \bar{w}_2)$ for $\gamma = 0.4$

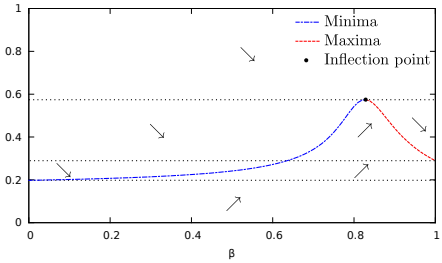


Fig. 7. Annotated version of Fig. 5a

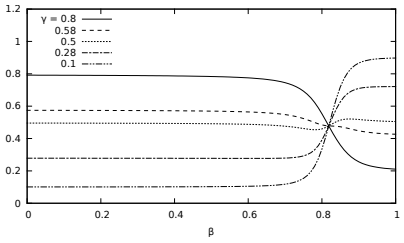


Fig. 8. Simulation of F_4 for several γ

intersects with $\phi(\beta)$. At the intersection point, the sign will change and the objective function will have an extremum or an inflection point. We can thus conclude that from a plot of $\phi^*(\bar{w}_2)$, we can determine the complete behaviour of the objective function without running any simulation or relying on possibly inaccurate approximate expressions for the performance characteristics.

In Fig. 8, finally, we have graphs of $F_4(\gamma, \beta)$, for several γ; they are obtained through simulation (see further for more details). We have chosen a γ-value from each area in Fig. 7 and we see that the graphs in Fig. 8 confirm our analysis. For $\gamma = 0.5$, for instance, we see that $F_4(\gamma, \beta)$ first reaches a minimum and later a maximum; for $\gamma = 0.58$, $F_4(\gamma, \beta)$ has an inflection point. We see from this study that the behaviour of the objective function largely depends on the coefficient γ in that function. Using our theorems presented here, these different behaviours can be seen at a glance from a graph of $\phi^*(\bar{w}_2)$.

Let us now return to the sensitivity analysis. From Fig. 5b, an analist can see what the impact of a variation in γ will be on $\bar{w}_2$ (and subsequently also on the value of the objective function). An annotated version as presented in Fig. 7 easily shows how the objective function behaves for different γ. This figure can be used to make a selection of different values of γ for which the objective function can be studied more closely, as was done in Fig. 8. The mapping from $\bar{w}_2$ to β, however, is still unknown. To get this information, one needs to resort to, e.g., simulation. In the next section, we show how this can be done efficiently.

5 Achieving a Specific Performance Vector

In this section, we combine the results obtained in the previous sections to optimize the simulation and optimization procedure. We will optimize objective

functions of the form of (7). We do this by optimizing for 101 values of γ equally spaced in $[0, 1]$. We compare several techniques and see how they influence the simulation effort and execution time. Lastly, we give some closing remarks on how the procedure could be sped up even further. If you only need to optimize a simple objective function (without changing the γ parameter) the optimization is even faster, though completely analog to the one presented here.

For all simulations, we used one slot service times and the arrival process presented in Sect. 3. We used Monte-Carlo simulations over 10^8 slots. This length of trajectory is long enough to minimize bias and variance from the transient behaviour (before reaching steady state) and the selection of the specific trajectory. To guarantee, however, that Property 2 is fulfilled we use identical arrival and decision variable trajectories, i.e., in each M-th slot of every simulation the same number of packets arrives and the same decision variable to choose a queue to serve (which then needs to be compared to β) is used. We achieve this by initializing the random generator with the same seed for each simulation. This is the well known method of common random numbers (CRN) [1, 11].

For the numerical results in this section, we optimized the objective function F_3. During the simulation runs for different γ, we only do one simulation for a given value of β. A table in memory holds the already simulated values for β and their results. This is a first way to speed up the process. As we use the CRN-method, each simulation for the same weight β will give us the same result.

A first, albeit naive, method is to just simulate equally spaced β's in $[0, 1]$. We call this the *brute force method*. In a second method, we use a golden section search on the objective function, like we used in [13]. This method, however, is only usable for objective functions known to have a single extremum and this extremum being a minimum. This is, for instance, the case for F_3 with convex g_j.

A third method follows from Sect. 2. We know that it is not needed to work directly with the objective function. From the objective function, the optimal value of $\bar{w}_2$ can be calculated. Subsequently, the corresponding value of β to achieve this value of $\bar{w}_2$ needs to be obtained by simulation. Knowing that $\bar{w}_2(\beta)$ is monotonically increasing, we can use a simplified version of the golden section search method (mentioned in the previous paragraph). In this method, the algorithm maintains an interval $[A, B]$ for β (starting from $[0, 1]$) wherein the optimal solution can be found. Each iteration, the algorithm chooses a point C in the interval (according to the golden ratio). This point is subsequently simulated whereafter the algorithm updates the interval to either $[A, C]$ or $[C, B]$.

We can use two different stopping criteria. The first one is the β-precision. This is the value $B - A$; if this value is small enough, we stop the algorithm and use the average $(A + B)/2$ as value for β_{opt}. Another option, called the F-precision, is to stop the simulation once we found a β that leads to a *reasonable* value of F. This F-precision is the percentage deviation from the minimum of the objective function F (which we can calculate in advance, as shown in Sect. 2).

The computational effort of the different simulation methods for this specific case can be found in Table 1. Cases can be engineered where the efficiency order of these methods is different; however, these cases are exceptional and need to be tailor-made. Furthermore, golden section search on F has a limited usability as it can only be used for objective functions with a single extremum, specifically

Table 1. Simulation times

Method	Stopping criterium	Needed simulations
Brute force	β-precision: 0.0001	10001
Golden section on F	β-precision: 0.0001	897
Golden section on $\bar{w}_2$	β-precision: 0.0001	617
Golden section on $\bar{w}_2$	β-precision: 0.0001 or F-precision: 1%	12

a minimum. It is clear that in general the more information and knowledge you have about the queueing system and objective function, the less simulations are needed. This often leads to complex algorithms that are only usable in a limited number of cases. The framework presented here, being as simple and general as it is, leads to large simulation gains without significant increase in complexity.

The methods presented here can be improved even further. Instead of using the golden section search in its purest form, the table with the already simulated β's could be used to select a starting interval $[A, B]$ after which golden section search could be used to further refine the result. This would lead to faster convergence. Another method (variation on golden section search), is to use multiple cores and select multiple C's. This way the interval $[A, B]$ will shrink much faster. Lastly, one could also vary the number of simulated slots as we get further into the algorithm, simulating less slots (and having a rougher estimation) when the interval $[A, B]$ is still large. A word of caution however is in order here, as this also induces extra variance. This variance could cause the algorithm to exclude the minimum. Using the presented insights the algorithm can easily detect when this happens and act accordingly.

6 Conclusions

In this paper, we have shown that no simulations or complex approximation techniques are needed for two-class GPS to study a generic objective function in terms of mean performance characteristics. With our framework, we are able to calculate the number of local extrema and the values of the objective function in these extrema. In this way, we are able to characterize the entire behaviour of the objective function w.r.t. the GPS weights. Our framework is based on results of strict priority scheduling and some specific properties of two-class GPS. To find the weights that optimize GPS, we still need to resort to a simulation approach. However, knowing the behaviour of the objective function beforehand can aid immensely in this optimization process.

Acknowledgement

This research has been co-funded by the Interuniversity Attraction Poles (IAP) Programme initiated by the Belgian Science Policy Office.

References

1. Asmussen, S., Glynn, P.W.: Stochastic Simulation: Algorithms and Analysis, vol. 57. Springer (2007)
2. Bruneel, H., Kim, B.G.: Discrete-time models for communication systems including ATM. Kluwer Academic Publishers (1992)
3. Demoor, T., Walraevens, J., Fiems, D., Bruneel, H.: Performance analysis of a priority queue: Expedited forwarding PHB in diffserv. AEU-International Journal of Electronics and Communications 65(3), 190–197 (2011)
4. Fiedler, M., Hossfeld, T., Tran-Gia, P.: A generic quantitative relationship between quality of experience and quality of service. Network, IEEE 24(2), 36–41 (2010)
5. Homg, M.F., Lee, W.T., Lee, K.R., Kuo, Y.H.: An adaptive approach to weighted fair queue with QoS enhanced on IP network. In: TENCON 2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology. vol. 1, pp. 181–186. IEEE (2001)
6. Huang, T.Y.: Analysis and modeling of a threshold based priority queueing system. Computer Communications 24(3), 284–291 (2001)
7. Lee, H.W., Kim, C., Chong, S.: Scheduling and source control with average queue-length control in cellular networks. In: Communications, 2007. ICC'07. IEEE International Conference on. pp. 109–114. IEEE (2007)
8. Neely, M.J.: Delay-based network utility maximization. IEEE/ACM Transactions on Networking (TON) 21(1), 41–54 (2013)
9. Parekh, A.K., Gallager, R.G.: A generalized processor sharing approach to flow control in integrated services networks: the single-node case. IEEE/ACM Transactions on Networking (TON) 1(3), 344–357 (1993)
10. Parekh, A.K., Gallagher, R.G.: A generalized processor sharing approach to flow control in integrated services networks: the multiple node case. IEEE/ACM Transactions on Networking (TON) 2(2), 137–150 (1994)
11. Spall, J.C.: Introduction to stochastic search and optimization: estimation, simulation, and control, vol. 65. John Wiley & Sons (2005)
12. Van Moorsel, A.: Metrics for the internet age: Quality of experience and quality of business. In: Fifth International Workshop on Performability Modeling of Computer and Communication Systems, Arbeitsberichte des Instituts für Informatik, Universität Erlangen-Nürnberg, Germany. vol. 34, pp. 26–31 (2001)
13. Vanlerberghe, J., Maertens, T., Walraevens, J., De Vuyst, S., Bruneel, H.: A hybrid analytical/simulation optimization of generalized processor sharing. In: Proceedings of The 25th International Teletraffic Congress (ITC 25). Shanghai (September 2013)
14. Verloop, I.M., Ayesta, U., Borst, S.: Monotonicity properties for multi-class queueing systems. Discrete Event Dynamic Systems 20(4), 473–509 (2010)
15. Walraevens, J., Steyaert, B., Bruneel, H.: Performance analysis of a single-server ATM queue with a priority scheduling. Computers & Operations Research 30(12), 1807–1829 (2003)
16. Walraevens, J., Steyaert, B., Bruneel, H.: Performance analysis of a GI-Geo-1 buffer with a preemptive resume priority scheduling discipline. European Journal of Operational Research 157(1), 130–151 (2004)
17. Walraevens, J., Steyaert, B., Bruneel, H.: Analysis of a discrete-time preemptive resume priority buffer. European Journal of Operational Research 186(1), 182–201 (2008)
18. Walraevens, J., Vanlerberghe, J., Maertens, T., De Vuyst, S., Bruneel, H.: Strict monotonicity and continuity of mean unfinished work in two queues sharing a processor (forthcoming)